(19)日本国特許庁(JP)　(12)特　許　公　報(B2)　(11)特許番号

特許第6285943号
(P6285943)

(45)発行日　平成30年2月28日(2018.2.28)　(24)登録日　平成30年2月9日(2018.2.9)

(51)Int.Cl.		F I		
G06F 17/22	(2006.01)	G06F	17/22	682
G06N 99/00	(2010.01)	G06N	99/00	150
G06F 8/65	(2018.01)	G06F	9/06	630A

請求項の数 19　(全 20 頁)

(21)出願番号	特願2015-537832(P2015-537832)
(86)(22)出願日	平成25年10月17日(2013.10.17)
(65)公表番号	特表2016-502701(P2016-502701A)
(43)公表日	平成28年1月28日(2016.1.28)
(86)国際出願番号	PCT/US2013/065497
(87)国際公開番号	WO2014/062948
(87)国際公開日	平成26年4月24日(2014.4.24)
審査請求日	平成28年9月14日(2016.9.14)
(31)優先権主張番号	13/653,581
(32)優先日	平成24年10月17日(2012.10.17)
(33)優先権主張国	米国(US)

(73)特許権者　314015767
マイクロソフト　テクノロジー　ライセンシング，エルエルシー
アメリカ合衆国　ワシントン州　98052　レッドモンド　ワン　マイクロソフト　ウェイ
(74)代理人　100079108
弁理士　稲葉　良幸
(74)代理人　100109346
弁理士　大貫　敏史
(74)代理人　100117189
弁理士　江口　昭彦
(74)代理人　100134120
弁理士　内藤　和彦

最終頁に続く

(54)【発明の名称】文字列変換の帰納的合成のための順位付け

(57)【特許請求の範囲】

【請求項１】

各々が部分表現を含むプログラム表現を含む候補変換プログラムを順位付けして、ユーザーによって入力される入力文字列から、１つまたは複数のユーザー供給される入力－出力例の各々と調和する、各々がユーザー所望の形式で出力文字列を生成する１つまたは複数の変換プログラムの順位付けされたグループを確立するために、コンピューターにより実装されるプロセスであって、

コンピューターを使用して、以下のプロセス動作、すなわち、

前記１つまたは複数のユーザー供給される入力－出力例から帰納的に合成された、各々が各ユーザー供給される入力例から各ユーザー供給される出力例によって提示される形式で出力文字列を生成する候補変換プログラムのセットを入力することと、

各候補変換プログラムについて、

より小さい方からより大きい方への順に、順位付け体系が確立された、前記候補変換プログラムの各部分表現について、前記部分表現についての前記順位付け体系を使用して尤度スコアを確立することと、

前記候補変換プログラムについて確立された前記部分表現の尤度スコアから、前記候補変換プログラムについての全体的な順位付けスコアを算出することとを実行させることを含む、コンピューターにより実装されるプロセス。

【請求項２】

部分表現が、正規表現または位置表現またはアトミック表現または連結表現のいずれか

であり、前記順位付け体系が、前記正規表現、位置表現、アトミック表現、および連結表現の各々をこの順序で順位付けするために確立される、請求項１に記載のプロセス。
【請求項３】
アトミック表現が、定数文字列表現または部分文字列表現のいずれかであり、アトミック表現についての順位付け体系を確立するプロセス動作が、
前記アトミック表現が定数文字列表現であるのか、または部分文字列表現であるのかを識別する動作と、
各タスクが複数の入力－出力例から成るタスクのトレーニングセットを入力する動作と、
前記トレーニングセットからトレーニングデータを生成する動作と、
各部分文字列表現および各定数文字列表現に尤度スコアを割り当てるために、機械学習分類子をトレーニングする動作と含む、請求項２に記載のプロセス。
【請求項４】
各部分文字列表現に尤度スコアを割り当てるプロセス動作が、０から１までの範囲における値を有する尤度スコアを割り当てることを含み、各定数文字列表現に尤度スコアを割り当てるプロセス動作が、所定の尤度スコア値を割り当てることを含む、請求項３に記載のプロセス。
【請求項５】
前記タスクのトレーニングセットからトレーニングデータを生成するプロセス動作が、
前記トレーニングセット内の各タスクについて、
前記タスク内の入力－出力例ごとに、
前記入力－出力例から帰納的に合成され、各々が前記例内の入力文字列のタプルから前記例内の前記出力文字列を生成する、変換プログラムのセットを確立する動作と、
前記変換プログラムのセットから正のトレーニング部分表現のセットを確立する動作と、
前記変換プログラムのセットから負のトレーニング部分表現のセットを確立する動作と、
前記正および負のトレーニング部分表現のセット内の各部分文字列表現から前記トレーニングデータを生成する動作とを含む、請求項３に記載のプロセス。
【請求項６】
前記正および負のトレーニング部分表現のセット内の各部分文字列表現からトレーニングデータを生成する前記プロセス動作が、
特徴の各々が、前記出力文字列内の部分文字列が前記正および負のトレーニング部分表現のセット内の部分文字列表現によって生成されるかを示す、１つまたは複数の特徴のセットを入力する動作と、
前記正および負のトレーニング部分表現のセット内の各部分文字列表現について、各要素が、クラスラベルのために予約される要素を除いて、前記特徴のうちの１つにそれぞれ関連付けられる特徴値に対応する、要素の所定の順序を含む特徴ベクトルを生成する動作とを含み、前記特徴ベクトルの生成は、
前記要素の所定の順序に対応する順序における前記特徴の各々について、前記特徴が考慮中の前記部分文字列表現によって生成される前記出力文字列内の部分文字列において提示される場合、第１のバイナリ値を有する特徴値を、対応する特徴ベクトル要素に割り当て、前記特徴が考慮中の前記部分文字列表現によって生成される前記出力文字列中の部分文字列において提示されない場合、第２のバイナリ値を有する特徴値を、前記対応する特徴ベクトル要素に割り当てることと、
前記クラスラベルのために予約される前記特徴ベクトル要素について、前記部分文字列表現が前記正のトレーニング部分表現のセットに存在する場合、１つのバイナリ値を割り当て、前記部分文字列表現が前記負のトレーニング部分表現のセットに存在する場合、異なるバイナリ値を割り当てることとを含む、請求項５に記載のプロセス。
【請求項７】

前記１つまたは複数の特徴のセットは、
前記出力部分文字列の左端位置がトークンによって認識されること、
前記出力部分文字列の前記左端位置の文字が定数であること、
前記出力部分文字列の右端位置がトークンによって認識されること、
前記出力部分文字列の前記右端位置の文字が定数であること、
出力部分文字列はトークンを示すこと、
前記入力部分文字列の左端位置がトークンによって認識されること、
前記入力部分文字列の右端位置がトークンによって認識されること、
入力部分文字列はトークンであること、
前記出力部分文字列を取得するために実行されるケーシング、
前記部分文字列の長さ、
前記入力文字列の前記長さに対する前記部分文字列の相対的な長さ、
前記出力文字列の前記長さに対する前記部分文字列の相対的な長さ、または
構成要素となる位置表現の尤度スコア、
の少なくとも１つを含む、請求項６に記載のプロセス。

【請求項８】

前記機械学習分類子は、サポートベクターマシン（ＳＶＭ）バイナリ分類子である、請求項６に記載のプロセス。

【請求項９】

タスク内の入力－出力例から確立される前記変換プログラムのセットから正のトレーニング部分表現のセットを確立する前記プロセス動作が、前記タスク内の前記入力－出力例から合成される変換プログラムの全てのセットの共通部分において見出される各部分表現を、正のトレーニング部分表現として指定することを含む、請求項５に記載のプロセス。

【請求項１０】

前記タスク内の前記入力－出力例から確立される前記変換プログラムのセットから負のトレーニング部分表現のセットを確立する前記プロセス動作が、正のトレーニング部分表現として識別されない前記変換プログラムのセット内の各部分表現を、負のトレーニング部分表現として指定することを含む、請求項９に記載のプロセス。

【請求項１１】

正規表現および位置表現について順位付け体系を確立する前記プロセス動作が、有限な数の予め定義されたトークンを有するセットから値を取る出力部分文字列を生成する部分表現を、前記正のトレーニング部分表現のセット内のそれらの部分表現の頻度に基づいて、頻度に基づくスコアにマッピングする辞書を算出する動作を含む、請求項５に記載のプロセス。

【請求項１２】

連結表現について順位付け体系を確立する前記プロセス動作が、連結表現において見出される個々のアトミック表現についてのスコアを結合することを含む、請求項２に記載のプロセス。

【請求項１３】

連結表現において見出だされる個々のアトミック表現についてスコアを連結する前記プロセス動作が、前記スコアを乗算すること、または前記スコアを加算することの１つを含む、請求項１２に記載のプロセス。

【請求項１４】

１つまたは複数の変換プログラムの前記順位付けされたグループとして、所定の数の上位の候補変換プログラムを選択するプロセス動作をさらに含む、請求項１に記載のプロセス。

【請求項１５】

１つまたは複数の変換プログラムの前記順位付けされたグループとして、所定の数の上位の候補変換プログラムを選択する前記プロセス動作が、前記最高位の候補変換プログラムを選択することを含む、請求項１４に記載のプロセス。

【請求項１６】
各々のプログラムが部分表現を含むプログラム表現を含む候補変換プログラムを順位付けして、ユーザーによって入力される入力文字列から、１つまたは複数のユーザー供給される入力－出力例の各々と調和する、各々がユーザー所望の形式で出力文字列を生成する１つまたは複数の変換プログラムの順位付けされたグループを確立するためのシステムであって、
１つまたは複数のコンピューティング装置であって、前記コンピューティング装置は複数のコンピューティング装置がある場合はいつでもコンピューターネットワークを介して相互に通信する、１つまたは複数のコンピューティング装置と、
前記１つまたは複数のコンピューティング装置によって実行されるプログラムモジュールを有するコンピュータープログラムとを含み、前記１つまたは複数のコンピューティング装置は、
前記１つまたは複数のユーザー供給される入力－出力例から帰納的に合成された、各々が各ユーザー供給される入力例から各ユーザー供給される出力例によって提示される形式で出力文字列を生成する前記候補変換プログラムのセットを入力し、
各候補変換プログラムについて、
より小さい方からより大きい方への順に、順位付け体系が確立された、前記候補変換プログラムの各部分表現について、前記部分表現について確立された前記順位付け体系を使用して尤度スコアを確立し、
前記候補変換プログラムについて確立された前記部分表現の尤度スコアから、前記候補変換プログラムについての全体的な順位付けスコアを算出するように、前記コンピュータープログラムの前記プログラムモジュールにより誘導される、システム。
【請求項１７】
候補変換プログラムの順位付けの前に前記順位付け体系を確立するためのプログラムモジュールをさらに含み、各順位付け体系は前記部分表現が、ユーザーによって入力される入力文字列から、前記１つまたは複数のユーザー供給される入力－出力例の各々と調和する、ユーザー所望の形式で出力文字列を生成することができるプログラムの一部である尤度を示す尤度スコアを生成し、前記順位付け体系確立プログラムモジュールが、正規表現、位置表現、アトミック表現、および連結表現の各々をこの順序で順位付けする前記順位付け体系を確立するためのサブモジュールを含み、部分表現は、正規表現または位置表現またはアトミック表現または連結表現のいずれかである、請求項１６に記載のシステム。
【請求項１８】
順位付け体系が確立された前記候補変換プログラムの各部分表現について、前記部分表現のために確立された前記順位付け体系を使用して尤度スコアを確立する前記プログラムモジュールは、
正規表現および位置表現を頻度に基づくスコアにマッピングする辞書を採用することを含む、正規表現および位置表現のための頻度に基づく順位付けを使用することと、
アトミック表現のための特徴に基づく順位付けを使用することであって、アトミック表現は定数文字列表現または部分文字列表現のどちらかであり、前記特徴に基づく順位付けは、
前記アトミック表現が定数文字列表現または部分文字列表現のどちらかあるか特定することと、
前記アトミック表現が定数文字列表現である場合はいつでも、所定の尤度スコア値を割り当てることと、
前記アトミック表現が部分文字列表現である場合にはいつでも、尤度スコアを各部分文字列表現に割り当てるようにトレーニングされた機械学習分類子を使用して尤度スコア値を割り当てることと、を含む、アトミック表現のための特徴に基づく順位付けを使用することと、
連結表現において見出される個々のアトミック表現についてのスコアを結合することを含む、連結表現を順位付けるためのパスに基づく順位付けを使用すること、のためのサ

ブモジュールを含む、請求項１７に記載のシステム。
【請求項１９】
１つまたは複数の変換プログラムの前記順位付けされたグループとして、所定の数の前記上位の候補変換プログラムを選択するためのプログラムモジュールをさらに含む、請求項１６に記載のシステム。
【発明の詳細な説明】
【背景技術】
【０００１】
[001] 世界中の数百万人もの人々が、データの記憶および操作のためにスプレッドシート等を使用する。これらのデータ操作シナリオは、大量の入力情報を１つの形式から別の形式へ転換して、所望の出力を生成することに関与することが多い。典型的には、これらのタスクは、手動で、または、エンドユーザーもしくはエンドユーザーのためのプログラマのいずれかによって作成される、小さい、しばしば１回限りのコンピュータープログラムを使用して達成される。
【０００２】
[002] 別のアプローチは、コンピューターを採用して、所望のデータ変換を達成するためのプログラムを合成しようとする試みに関与した。プログラムを合成する２つの主要なアプローチ、すなわち、演繹法および帰納法が存在する。演繹的プログラム合成において、完全な上位の仕様は、翻訳の各工程が何らかの公理を使用して有効にされる、対応する下位のプログラムに翻訳される。このアプローチは、ユーザーが完全な仕様を提供することを必要とし、これは、幾つかの場合において、プログラム自体を書くことよりも潜在的に困難となり得る。これは、帰納的合成アプローチを、近年より普及させることとなった。帰納的プログラム合成アプローチにおいて、プログラムは、入力－出力例のセットから成る仕様などの不完全な仕様から合成される。帰納的プログラム合成アプローチは、下位のポインター操作コードからスプレッドシートのマクロに及ぶ、様々なドメインからのプログラムを合成するために近年使用されてきた。
【０００３】
[003] 帰納的プログラム合成アプローチにおける仕様は、不完全であり、しばしば曖昧であるため、所与の仕様と調和する、基になるドメイン固有言語の異なるプログラムが存在する。曖昧さを除去し、所望のプログラムへ収束させるために、ユーザーは、付加的な入力－出力例を提供することによって、仕様を強化する必要がある。例の数は、ドメイン固有言語の表現度と正比例する。すなわち、言語が表現的であるほど、所望のプログラムに収束させるためにより多くの入力－出力例が必要とされる。
【０００４】
[004] ドメイン固有言語は、ユーザーが望む大抵のタスクを表現するために表現的である必要があるが、同時に、所望のプログラムを学習するための面倒な数の入力－出力例をユーザーが提供することは期待できない。
【発明の概要】
【０００５】
[005] 本明細書において説明される順位付け技法の実施形態は、一般に、帰納的プログラム合成手続を通じて生成された候補変換プログラムを、少数のユーザー供給された入力－出力例を使用して、順位付けすることに関与する。一実施形態において、コンピューターは、候補変換プログラムを順位付けして、１つまたは複数の変換プログラムの順位付けされたグループを確立するために使用される。１つまたは複数の変換プログラムの各々は、ユーザーによって入力された各入力文字列から、１つまたは複数のユーザー供給された入力－出力例の各々に調和する、ユーザー所望の形式における出力文字列を生成する。
【０００６】
[006] より詳細には、候補変換プログラムのセットが入力される。各変換プログラムは、ドメイン固有言語のプログラム表現から構成され、プログラム表現は、部分表現から構成されることに留意されたい。順位付け体系は、候補変換プログラムにおいて見出される

部分表現についてのオフライントレーニングデータから確立される。これらの順位付け体系の各々は、部分表現が、ユーザーによって入力された各入力文字列から、ユーザー供給される入力－出力例の各々と調和する、ユーザー所望の形式で出力文字列を生成することが可能なプログラムの一部である尤度を示す尤度スコアを生成する。順位付け体系が確立された各候補変換プログラムの各部分表現について、尤度スコアは、その部分表現について確立された順位付け体系を使用して確立される。尤度スコアは、より小さい部分表現からより大きい部分表現の順序で算出される。次いで、候補変換プログラムごとに、その候補変換プログラムについて確立された部分表現尤度スコアから、全体的な順位付けスコアが算出される。

【０００７】

[007] 一実施形態において、候補変換プログラムの各々に関連付けられる全体的な順位付けは、所定の数の上位の候補変換プログラムを選択するために使用される。次いで、これらの上位のプログラムは、１つまたは複数の変換プログラムの前述された順位付けされたグループとして指定される。１つの実装において、所定の数は１であり、従って、最高位の候補変換プログラムのみが選択および指定されることに留意されたい。

【０００８】

[008] 本概要は、詳細な説明においてさらに後述される概念の集まりを簡略化された形式で紹介するために提供されることに留意すべきである。本概要は、特許請求の範囲に記載された主題の重要な特徴または本質的な特徴を識別することを意図されたものではなく、特許請求の範囲に記載された主題の範囲を判定する補助として使用されることを意図されたものでもない。

【０００９】

[009] 本開示の具体的な特徴、態様、および利点は、以下の説明、添付の特許請求の範囲、および付属の図面に関して、より良く理解されるであろう。

【図面の簡単な説明】

【００１０】

【図１】[0010]候補変換プログラムを順位付けするためのプロセスの一実施形態を一般的に概説するフロー図である。

【図２】[0011]アトミック表現のための順位付け体系を確立することに関与する、図１のプロセスの一部の実装を一般的に概説するフロー図である。

【図３】[0012]タスクのトレーニングセットからトレーニングデータを生成することに関与する、図２のプロセスの一部の実装を一般的に概説するフロー図である。

【図４】[0013]正および負のトレーニング部分表現のセット内の各部分文字列からトレーニングデータを生成することに関与する、図３のプロセスの一部の実装を一般的に概説するフロー図である。

【図５】[0014]順位付け体系が確立された候補変換プログラムの各部分表現についての尤度スコアを、その順位付け体系を使用して確立することに関与する、図１のプロセスの一部の実装を一般的に概説するフロー図である。

【図６】[0015]最高位の候補変換プログラムを使用して、文字列変換を実行するためのプロセスの一実施形態を一般的に概説するフロー図である。

【図７】[0016]人の名字を表す入力文字列を含む入力列と、文字列「Ｍｒ．」が名字の前に配置される対応する入力文字列名を含む１つの例示的な出力文字列を有する出力列とを有するスプレッドシートを描く表である。

【図８】[0017]都市名を含む住所を含む入力列と、都市名に対応する、対応する入力文字列の一部を含む１つの例示的な出力文字列を有する出力列とを有するスプレッドシートを描く表である。

【図９】[0018]本明細書において説明される順位付け技法の実施形態を実装するための例示的なシステムを構成する汎用コンピューティング装置を描く図である。

【発明を実施するための形態】

【００１１】

[0019] 順位付け技法の実施形態の以下の説明において、この説明の一部を形成し、本技法が実施され得る具体的な実施形態が例示として示される付属の図面への参照がなされる。本技法の範囲から逸脱することなく、他の実施形態が利用され、構造的な変更が行われ得ることが理解される。

【００１２】

[0020] 明確さのために、本明細書において説明される順位付け技法の実施形態を説明する際には具体的な専門用語が行使され、これらの実施形態がそのように選ばれた具体的な用語に限定されることは意図されないことにも留意されたい。さらに、各具体的な用語は、同様の目的を実現するために概ね同様に動作する、その用語のあらゆる技術的な均等物を含むことが理解されるべきである。本明細書における、「一実施形態」、もしくは「別の実施形態」、もしくは「例示的な実施形態」、もしくは「代替的な実施形態」、または「１つの実装」、もしくは「別の実装」、もしくは「例示的な実装」、もしくは「代替的な実装」への言及は、その実施形態または実装に関連して説明される特定の特徴、特定の構造、または特定の特性が、本順位付け技法の少なくとも１つの実施形態に含まれ得ることを意味する。本明細書中の様々な箇所における「一実施形態において」、「別の実施形態において」、「例示的な実施形態において」、「代替的な実施形態において」、「１つの実装において」、「別の実装において」、「例示的な実装において」、「代替的な実装において」という表現の出現は、必ずしも全てが同じ実施形態または実装に言及しているわけではなく、別個または代替的な実施形態／実装は、他の実施形態／実装と相互に排他的でもない。またさらに、本順位付け技法の１つまたは複数の実施形態または実装を表すプロセスフローの順序は、いかなる特定の順序も本質的に示さず、本順位付け技法のいかなる限定も示唆しない。

【００１３】

[0021] 本明細書において使用される「入力－出力例」という用語は、入力を形成する文字列のタプルおよび出力を形成する文字列に言及することにさらに留意されたい。入力は、ユーザーが変換することを希望する入力の例を表すのに対して、出力文字列は、ユーザーが入力から生成されることを希望する出力の例を表す。

【００１４】

１．０ 帰納的プログラム合成のための順位付け技法

[0022] 本明細書において説明される順位付け技法の実施形態は、統計的および機械的学習技法を使用して、帰納的プログラム合成において使用するための所望の順位付け関数を学習する。一般に、正および負の例のトレーニングデータセットは、入力－出力例の所与のセットから自動的に作成される。トレーニングデータセットから、ドメイン固有言語のプログラムにおける表現を尤度測度に割り当てる順位付け関数が学習される。次いで、これらの順位付け関数が使用されて、ごく少数の入力－出力例から、学習されたプログラムの尤度が算出される。

【００１５】

１．１ 文字列変換言語L_S

[0023] このセクションでは、順位付け技法の実施形態を実装するための文字列変換言語が説明される。文字列変換言語L_Sについての構文は、以下の通りである。

Ｔｒａｃｅ ｅｘｐｒ ｅ ：＝Ｃｏｎｃａｔｅｎａｔｅ（f_1，．．．，f_n）
Ａｔｏｍｉｃ ｅｘｐｒ ｆ ：＝ＳｕｂＳｔｒ（v_i，p_1，p_2）
｜ＣｏｎｓｔＳｔｒ（ｓ）
Ｐｏｓｉｔｉｏｎ ｅｘｐｒ ｐ ：＝Ｃｐｏｓ（ｋ）｜Ｐｏｓ（r_1，r_2，ｃ）
Ｉｎｔｅｇｅｒ ｅｘｐｒ ｃ ：＝ｋ
Ｒｅｇｕｌａｒ ｅｘｐｒ ｒ ：＝ＴｏｋｅｎＳｅｑ（T_1，．．．，T_m）．

【００１６】

[0024] トレース（または連結）表現ｅは、アトミック表現f_1，．．．，f_nの連結を示す。アトミック表現ｆは、定数文字列表現ＣｏｎｓｔＳｔｒ（ｓ）または部分文字列表現ＳｕｂＳｔｒ（v_i，p_1，p_2）のいずれかを示し得る。定数文字列表現Ｃｏｎｓｔ

Ｓｔｒ（ｓ）は、定数文字列「ｓ」を生成する表現を示す。部分文字列表現ＳｕｂＳｔｒ（$ｖ_{ｉ}$，$ｐ_{1}$，$ｐ_{2}$）は、列$ｖ_{ｉ}$に存在し、その左端位置および右端位置がそれぞれ位置ペア$ｐ_{1}$および$ｐ_{2}$によって表される入力文字列の部分文字列を表す。位置表現Ｃｐｏｓ（ｋ）は、整数定数が正（または負）である場合、左側（または右側）からの所与の文字列中のｋ番目のインデックスを指す。位置表現Ｐｏｓ（$ｒ_{1}$，$ｒ_{2}$，ｃ）は、その左側および右側がそれぞれ正規表現$ｒ_{1}$および$ｒ_{2}$に一致する入力文字列中の位置を示し、その文字列における正規表現のｃ番目のそのような一致である。正式には、正規表現$ｒ_{1}$は、ｓ［０．．ｔ－１］の何らかのサフィックスに一致し、$ｒ_{2}$は、ｓ［０．．ｌ＋１］の何らかのサフィックスに一致し、ただし、ｌ＝Ｌｅｎｇｔｈ（ｓ）であり、ｔは、ｃが正（または負）である場合、左側（または右側）から開始するｃ番目のそのようなインデックス／位置である。

【００１７】

１．２　$Ｌ_{Ｓ}$表現の順位付けセット

[0025]　$Ｌ_{Ｓ}$表現の大きなセットを表すデータ構造は、以下の通りである。

【００１８】

【数１】

$$\tilde{e} \;:=\; \mathcal{D}ag\left(\tilde{\eta}, \eta^{s}, \eta^{t}, \tilde{\xi}, W\right)$$

$$\tilde{f} \;:=\; \mathrm{SubStr}\left(v_i, \{p_j\}_j, \{p_k\}_k\right) \;\Big|\; \mathrm{ConstStr}(s)$$

$$p \;:=\; \mathrm{CPos}(k) \mid \mathrm{Pos}(r_1, r_2, c)$$

$$r \;:=\; \mathrm{TokenSeq}(T_1, .., T_m)$$

【００１９】

[0026]　データ構造は、多項式空間における表現の指数部つきの数値を表すために、ある表現の独立した部分表現を独立して維持する。例えば、部分文字列のセットＳｕｂＳｔｒ（$ｖ_{ｉ}$，$\{ｐ_{ｊ}\}_{ｊ}$，$\{ｐ_{ｋ}\}_{ｋ}$）は、左端位置の式（$\{ｐ_{ｊ}\}_{ｊ}$）および右端位置の式（$\{ｐ_{ｋ}\}_{ｋ}$）を独立して維持する。マッピングＷは、ξにおける各エッジをＳｕｂＳｔｒおよびＣｏｎｓｔＳｔｒのセットにマッピングし、それによって、異なる長さの連結表現間での共有さえ実現する（開始ノードη^{S}から対象ノードη^{t}までの無閉路有向グラフにおける各パスは、そのパスにおける各エッジからの任意の表現を取り出して、それらを連結することによって取得された連結表現のセットを表す）。本明細書において説明される順位付け技法の実施形態は、この部分表現の独立性を維持し、部分表現のセットを独立して順位付けする。部分表現のセットを順位付けするための３つの技法、すなわち、（正規表現ｒおよび位置表現ｐを順位付けするための）頻度に基づく順位付け、（Ｓｕｂｓｔｒ表現を順位付けするための）特徴に基づく順位付け、および（連結表現を順位付けするための）パスに基づく順位付けが使用される。

【００２０】

１．２．１　頻度に基づく順位付け

[0027]　正規表現および位置表現（これらの双方は、何らかの有限のセットから値を取る）の場合、そのような表現を順位付けするために、頻度に基づく順位付けが実行される。

取り得る各表現値をトレーニングデータから推定されるその表現値の頻度にマッピングする辞書Ｄが存在する。正規表現の尤度スコアは、その頻度スコアに正比例するように定義される。位置表現の尤度スコアは、その頻度スコア、およびその構成要素である正規表現の尤度スコアの平方根の任意の線形関数となるように定義され得る。

【００２１】

１．２．２　特徴に基づく順位付け

[0028]　無限のセットから値を取るアトミック表現ｆなどの表現の場合、特徴に基づく順位付けが実行される。特徴のセットφ（ｅ）は、基となる入力－出力例から算出され、尤度は、機械学習手続から取得される。アトミック表現ｆについての定数文字列表現と部分文字列表現との間の選択を順位付けするために、サポートベクターマシン（SVM:support vector machine）バイナリ分類手続が、一実施形態において使用される。

【００２２】

１．２．３　パスに基づく順位付け

[0029]　アトミック表現から構成される連結表現の場合、パスに基づく順位付けが実行される。一実施形態において、このパスに基づく順位付けは、連結表現において見出される個々のアトミック表現についてのスコアを結合することに関与する。１つの実装において、アトミック表現についての尤度スコアは、これらを乗算することによって結合され、別の実装において、これらは互いに加算される。

【００２３】

１．３　自動化されたトレーニングデータ生成

[0030]　このセクションでは、頻度に基づく順位付け（または、それぞれ特徴に基づく順位付け）についての入力－出力例のセットから、正の例（または、正および負の例）から成るトレーニングデータを自動的に生成するための方法の一実施形態が説明される。各々が複数の入力－出力例｛（i_1，o_1），．．．，（i_n，o_n）｝から成るタスクの大きなセットが取得される。合成手続は、各入力－出力例（i_k，o_k）と調和する全てのプログラム

【００２４】

【数２】

$$\tilde{P}_k$$

のセットを学習する。次いで、正および負の表現値は、以下のように取得される。正の表現値は、セット

【００２５】

【数３】

$$\tilde{P}_1 \cap \tilde{P}_2 \ldots \cap \tilde{P}_n$$

に存在する表現値から構成され、負の表現値は、セット

【００２６】

【数４】

$$\left\{\tilde{P}_k \setminus \left(\tilde{P}_1 \cap \tilde{P}_2 \ldots \cap \tilde{P}_n\right) \mid 1 \leq k \leq n\right\}$$

に存在する表現値から構成される。

【００２７】

１．３．１　正および負の表現値

[0031]　２つの無閉路有向グラフ（dags）D_kおよび$D_\cap \equiv (D_1 \cap D_2 . . . D_n)$が与えられると、課題は、２つのｄａｇにおけるエッジを揃えて、正および負の表現値を算出することである。ｄａｇを揃えた後、これらの間の共通エッジは、正の表現値を構成し、D_kには存在するが、$D_\cap$には存在しないエッジは、負の表現値を構成する。

【００２８】
[0032] ＤＡＧプログラムD_kおよび$D_\cap$は、入力文字列i_kに対して実行され、ｄａｇノードは、ラベル関数Ｌ：η→ｉｎｔを使用して、出力文字列o_kのインデックスを用いて注釈付けされる。ｄａｇの開始ノードη^Sは、Ｌ（η^S）＝０となるように、インデックス０を用いて注釈付けされる。Ｌ（η_1）＝ｌであり、かつ、ｄａｇエッジ上の表現（η_1，η_2）が、入力文字列i_kに対して実行される場合において、文字列o_kを生成するとき、ｄａｇにおけるノードη_2は、インデックスｍ（すなわち、Ｌ（η^2）＝ｍ）を用いて注釈付けされる。双方のｄａｇのノードが注釈付けされると、同じラベルを有するノード間のエッジ上の表現は、比較のために収集される。ｄａｇD_kにおいて文字列o_k［１．．ｍ］を生成する表現のセットは、

【００２９】
【数５】
$$\tilde{e}_{l,m,k}$$
と示され、ただし、

【００３０】
【数６】
$$\tilde{e}_{l,m,k} \equiv \bigcup_{\eta_1,\eta_2,\in \mathcal{D}_k,} \tilde{e}(\eta_1,\eta_2), \mathrm{L}(\eta_1) = l, \mathrm{L}(\eta_2) = m$$
である。

【００３１】
【数７】
$$\tilde{e}_{l,m,\wedge}$$
において出現する表現は、正の表現として示され、セット

【００３２】
【数８】
$$\tilde{e}_{l,m,k} \setminus \tilde{e}_{l,m,\wedge}$$
において出現する表現は、負の表現として示される。正および負の表現値のセットは、各入力－出力例のペアについて取得される。

【００３３】
１．３．２　頻度に基づくトレーニングデータ生成
[0033] 正規表現および位置表現を順位付けする場合、これらの双方は有限のセットから値を取るため、頻度に基づく順位付けアプローチが実行される。トークンシーケンス表現についてのトークンシーケンスとこれらの頻度スコアのデータベースも、作成される。頻度は、正のトレーニング表現のセットから推定される。頻度は、ある表現が発生する各異なるコンテキストについても推定され得る。例えば、ある位置表現は、左端位置表現または右端位置表現のどちらかであり得る（その位置表現がＳｕｂＳｔｒ表現の第１の引数として発生するか、または第２の引数として発生するかを意味する）。正規表現も、ある位置表現内の２つの異なるコンテキストにおいて発生する。

【００３４】
１．３．３　特徴に基づくトレーニングデータ生成
[0034] 文字列変換を学習する際の主要な曖昧さのうちの１つは、出力文字列中の部分文字列が定数文字列であるのか、または入力文字列の部分文字列であるのかに関する決定を行うことに由来する。そのような決定を行うことは、位置ペアおよび定数文字列についての取り得る値だけでなく、入力文字列および出力文字列にも依存するため、無限に多くの取り得る入力文字列および出力文字列が存在することから、この場合において頻度に基づく順位付けを使用することは不可能である。代わりに、一実施形態において、特徴に基づく順位付けアプローチが使用されて、ＳｕｂＳｔｒ表現とＣｏｎｓｔＳｔｒ表現との間で

選択がされる。より詳細には、特徴のセットが、各ＳｕｂＳｔｒ表現について定義される。使用され得る特徴の例は、以下を含むが、以下に限定されない。
ａ）ＩｓＯｕｔｐｕｔＬｅｆｔＴｏｋ：出力部分文字列の左端位置がトークンによって認識され得るかを示すブール値
ｂ）ＩｓＯｕｔｐｕｔＬｅｆｔＣｏｎｓｔａｎｔ：出力部分文字列の左端位置の文字が定数であるかを示すブール値
ｃ）ＩｓＯｕｔｐｕｔＲｉｇｈｔＴｏｋ：出力部分文字列の右端位置がトークンによって認識され得るかを示すブール値
ｄ）ＩｓＯｕｔｐｕｔＲｉｇｈｔＣｏｎｓｔａｎｔ：出力部分文字列の右端位置の文字が定数であるかを示すブール値
ｅ）ＩｓＯｕｔｐｕｔＴｏｋ：出力部分文字列がトークンを意味するかを意味するブール値
ｆ）ＩｓＩｎｐｕｔＬｅｆｔＴｏｋ：入力部分文字列の左端位置がトークンによって認識され得るかを意味するブール値
ｇ）ＩｓＩｎｐｕｔＲｉｇｈｔＴｏｋ：入力部分文字列の右端位置がトークンによって認識され得るかを意味するブール値
ｈ）ＩｓＩｎｐｕｔＴｏｋ：入力部分文字列がトークンであるかを意味するブール値
ｉ）Ｃａｓｉｎｇ：出力部分文字列を取得するために実行されるケーシング
ｊ）ＬｅｎＳｕｂｓｔｒｉｎｇ：部分文字列の長さ
ｋ）ＲｅｌＬｅｎＩｎＳｕｂｓｔｒｉｎｇ：入力文字列の長さに対する部分文字列の相対的な長さ（lenSubstring/lenInputString）
ｌ）ＲｅｌＬｅｎＯｕｔＳｕｂｓｔｒｉｎｇ：出力文字列の長さに対する部分文字列の相対的な長さ（lenSubstring/len0utputString）
ｍ）頻度に基づく順位付け体系を使用して推定される、構成要素となる位置表現の尤度スコア

【００３５】

[0035] これらの特徴全ては、一定のＯ（ｌ）時間で算出され得る。各正および負のＳｕｂＳｔｒ表現について、特徴ベクトルは、クラスラベル（例えば、正の表現について＋１、および負の表現について０）と共に算出される。次いで、既製のサポートベクターマシン（SVM）手続が、バイナリ分類子を学習するために使用され得る。より詳細には、作成された各特徴ベクトルは、各要素（クラスラベルのために予約される要素を除く）が前述された特徴のうちの１つにそれぞれ関連付けられる特徴値に対応する所定の順序を有する。要素の所定の順序に対応する順序における前述された特徴の各々について、その特徴が考慮中のＳｕｂＳｔｒ表現において提示される場合、第１のバイナリ値（例えば、１）を有する特徴値が、対応する特徴ベクトル要素に割り当てられる。その特徴が提示されない場合、第２のバイナリ値（例えば、０）を有する特徴値が、対応する特徴ベクトル要素に割り当てられる。また、前述されたクラスラベルのために予約される特徴ベクトル要素は、ＳｕｂＳｔｒ表現が正のＳｕｂＳｔｒ表現である場合、１つのバイナリ値（例えば、１）を割り当てられ、ＳｕｂＳｔｒ表現が負のＳｕｂＳｔｒ表現である場合、別のバイナリ値（例えば、０）を割り当てられる。

【００３６】

１．４　順位付けプログラム

[0036] このセクションでは、ｄａｇによって表されるプログラムがどのように順位付けされるかが説明されるであろう。

【００３７】

１．４．１　Ｄａｇエッジ表現の順位付け

[0037] ｄａｇの各エッジは、ＳｕｂＳｔｒ表現およびＣｏｎｓｔＳｔｒ表現のセットから成る。一実施形態において、特徴に基づく順位付けは、＋１と０との間の尤度スコアを各ＳｕｂＳｔｒ表現に割り当てる一方で、ＣｏｎｓｔＳｔｒ表現の尤度スコアは、０．５になるように決められる。

【００３８】
１．４．２　Ｄａｇパスの順位付け
[0038] ｄａｇＤにおける各パスは、所与の出力例に適合する幾つかのプログラムを表す。エッジｅの尤度スコアｗ（ｅ）は、そのエッジにおける任意のＳｕｂＳｔｒ表現またはＣｏｎｓｔＳｔｒ表現の最も高い尤度スコアとして定義される。パスの尤度スコアは、そのパス上の対応するエッジ表現の尤度スコアを乗算する（加算する）ことによって算出されるように、ｗ（ｐ）≡$\Pi_{e \in edges(p)}$ｗ（ｅ）として定義される。次いで、ダイクストラの最短パス手続が使用されて、ｄａｇにおける最高位のパスが算出され、最高位の出力文字列を生成するために、このパスが実行される。パスの尤度スコアは、全てのエッジの尤度スコアの積／和、エッジの数、任意のエッジの最小／最大の尤度スコアなど、様々な特徴のより洗練された関数とすることもできる。さらに、この関数は、機械的学習技法を使用することによって学習されることもできる。
【００３９】
１．５　候補変換プログラムを順位付けするための例示的なプロセス
[0039] 本明細書において説明される順位付け技法の実施形態の前述の態様は、１つの一般的な実装において、図１において概説されるプロセスによって実現され得る。より詳細には、（後述の例示的な動作環境において説明されるコンピューティング装置のいずれかなどの）コンピューターは、候補変換プログラムを順位付けして、１つまたは複数の変換プログラムの順位付けされたグループを確立するために使用される。１つまたは複数の変換プログラムの各々は、ユーザーによって入力された各入力文字列から、１つまたは複数のユーザー供給された入力－出力例の各々と調和する、ユーザー所望の形式における出力文字列を生成する。より詳細には、候補変換プログラムのセットは、入力として受信される（プロセス動作１００）。候補変換プログラムの各々は、１つまたは複数のユーザー供給された入力－出力例から、従来の方法を使用して帰納的に合成され、各ユーザー供給された入力例から各ユーザー供給された出力例によって提示される形式で出力文字列を生成する。既述したように、各変換プログラムは、部分表現から構成されるプログラム表現から構成されることに留意されたい。次に、（より小さい部分表現からより大きい部分表現まで）順位付け体系が確立された各候補変換プログラムの各部分表現について、尤度スコアは、その部分表現について確立された順位付け体系を使用して確立される（プロセス動作１０２）。順位付け体系の各々は、部分表現が、ユーザーによって入力された各入力から、ユーザー供給された入力－出力例の各々と調和する、ユーザー所望の形式で出力文字列を生成することが可能なプログラムの一部である尤度を示す尤度スコアを生成することに留意されたい。次いで、全体的な順位付けスコアが、候補変換プログラムごとに、その候補変換プログラムについて確立された部分表現尤度スコアから算出される（プロセス動作１０４）。
【００４０】
[0040] 色々な種類の順位付け体系が、色々な種類の表現について使用される。既述したように、部分表現は、正規表現または位置表現またはアトミック表現または連結表現のいずれかであり得る。一実施形態において、正規表現および位置表現は、頻度に基づく順位付けを採用する（このタイプの順位付けは、トークンの有限なセットからの値を取る表現に適用可能であるため）。また、特徴に基づく順位付けは、アトミック表現について使用され、パスに基づく順位付けは、連結について使用される。効率のために、一実施形態において、順位付け体系は、まず正規表現に、次いで位置表現に、次いでアトミック表現に、最後に連結表現について確立されることにも留意されたい。
【００４１】
[0041] アトミック表現の場合において、既に示されたように、これらは、定数文字列表現または部分文字列表現のどちらかであり得る。これを考慮して、図２を参照すると、一実施形態においてアトミック表現についての順位付け体系を確立することは、まず、アトミック表現が定数文字列表現であるのか、または部分文字列表現であるのかを識別することに関与する（プロセス動作２００）。次いで、タスクのトレーニングセットが入力され

る（プロセス動作２０２）。各タスクセットは、複数の入力－出力例から構成される。トレーニングデータは、トレーニングセットから生成され（プロセス動作２０４）、尤度スコアを各部分文字列表現および各定数文字列表現に割り当てるべく、機械学習分類子をトレーニングするために使用される（プロセス動作２０６）。一実施形態において、機械学習分類子は、サポートベクターマシン（SVM）バイナリ分類子であり、部分表現が意図されるプログラムの一部である可能性がどのくらいあるかに応じて０から１までの範囲における値を有する各部分文字列表現に尤度スコアを割り当てるためにトレーニングされる。また、分類子は、所定の尤度スコア値（例えば、０．５）を各定数文字列表現に割り当てる。

【００４２】

[0042] タスクのトレーニングセットからトレーニングデータを生成することに関して、１つの実装において、これは、トレーニングセット内の各タスクについて以下に関与する。図３を参照すると、タスクにおいて、以前に選択されていない入力－出力例が選択される（プロセス動作３００）。変換プログラムのセットは、選択された入力－出力例から帰納的に合成される（プロセス動作３０２）。変換プログラムの各々は、その例の入力における入力文字列のタプルから、その例の出力において提示される形式で出力文字列を生成する。入力文字列のタプルとは、入力－出力例の出力を生成するために使用される入力－出力例における入力の１つまたは複数の部分をいう。次に、正のトレーニング部分表現のセットが、変換プログラムのセットから確立され（プロセス動作３０４）、負のトレーニング部分表現のセットも、変換プログラムのセットから確立される（プロセス動作３０６）。一実施形態において、変換プログラムのセットから正のトレーニング部分表現を確立することは、タスク内の入力－出力例から合成される変換プログラムの全てのセットの共通部分において見出される各部分表現を、正のトレーニング部分表現に指定することに関与する。一方で、変換プログラムのセットから負のトレーニング部分表現を確立することは、正のトレーニング部分表現として識別されない変換プログラムのセット内の各部分表現を、負のトレーニング部分表現として指定することに関与する。

【００４３】

[0043] この時点で、正規表現および位置表現についての順位付け体系が確立される。より詳細には、有限な数の予め定義されたトークンを有するセットから値を取る出力部分文字列を生成する部分表現（すなわち、正規表現および位置表現）を、正のトレーニング部分表現のセット内のそれらの部分表現の頻度に基づいて、頻度に基づくスコアにマッピングする辞書が生成される（プロセス動作３０８）。

【００４４】

[0044] 次に、部分表現の正および負のトレーニングセットの双方からの各部分文字列表現について、トレーニングデータが、特徴に基づく順位付けのために生成される（プロセス動作３１０）。次いで、タスク内の全ての入力－出力例が選択および処理されたかが判定される（プロセス動作３１２）。タスク内の全ての入力－出力例が選択および処理されていない場合、プロセス動作３００～３１２が繰り返される。そうでない場合、本手続は、選択されたタスクについて終了する。前述の手続は、残りのタスク全ておよびそれらの入力－出力例について繰り返される。

【００４５】

[0045] 特徴に基づく順位付けのために、正および負のトレーニング部分表現のセット内の各部分文字列表現からトレーニングデータを生成する前述のプロセス動作に関して、これは、図４に示されるような一実施形態において達成される。まず、（既述された特徴などの）１つまたは複数の特徴のセットが入力される（プロセス動作４００）。特徴の各々は、出力文字列中の部分文字列が正および負のトレーニング部分表現のセット内の部分文字列表現のうちの１つによって生成されるかを示す。次いで、正および負のトレーニング部分表現のセット内の各部分文字列表現について、特徴ベクトルが生成される（プロセス動作４０２）。既述したように、各特徴ベクトルは、クラスラベルのために予約された要素を除いて、各要素が前述された特徴のうちの１つにそれぞれ関連付けられる特徴値に対

応する、要素の所定の順序である。一実施形態において、要素の所定の順序に対応する順序における特徴の各々について、その特徴が考慮中の部分文字列表現によって生成される出力文字列中の部分文字列において提示される場合、各特徴ベクトルは、第１のバイナリ値（例えば、１）を有する特徴値を割り当てることによって生成される。より具体的には、特徴値は、提示される特徴に関連付けられる特徴ベクトル要素に割り当てられる。また、第２のバイナリ値（例えば、０）を有する特徴値は、考慮中の特徴が部分文字列表現によって生成される出力文字列中の部分文字列において提示されない場合、その特徴に関連付けられる特徴ベクトル要素に割り当てられる。さらに、前述されたクラスラベルのために予約される特徴ベクトル要素は、部分文字列表現が正のトレーニング部分表現のセット内にある場合、１つのバイナリ値（例えば、１）を割り当て、部分文字列表現が負のトレーニング部分表現のセット内にある場合、異なるバイナリ値（例えば、０）を割り当てる。

【００４６】

[0046] 連結表現について順位付け体系を確立することに関して、一実施形態において、これは、連結表現において見出される個々のアトミック表現についての尤度スコアを結合することに関与する。１つの実装において、この結合は、スコアを乗算することによって達成され、かつ１つの実装では、この結合は、スコアを加算することによって達成される。

【００４７】

[0047] 前記を考慮すると、一実施形態において、順位付け体系を使用してその順位付け体系が確立された候補変換プログラムの各部分表現についての尤度スコアを確立することは、以下に関与する。図５を参照すると、まず、各正規表現についての尤度が、頻度に基づく順位付け体系を使用して識別される（プロセス動作５００）。次いで、各位置表現についての尤度も、頻度に基づく順位付け体系を使用して識別される（プロセス動作５０２）。既述したように、この頻度に基づく順位付けは、正規表現および位置表現を頻度に基づくスコアにマッピングする辞書を採用することに関与し得る。次に、各アトミック表現の尤度が、特徴に基づく順位付け体系を使用して識別される（プロセス動作５０４）。既述したように、アトミック表現は、定数文字列表現または部分文字列表現のどちらかである。アトミック表現が定数文字列表現である場合、特徴に基づく順位付けは、所定の尤度スコア値（例えば、０．５）を割り当てることに関与し得る。また、アトミック表現が部分文字列表現である場合、特徴に基づく順位付けは、尤度スコアを各部分文字列表現に割り当てるようにトレーニングされた機械学習分類子を使用して、（例えば、０から１までに及ぶ）尤度スコア値を割り当てることに関与し得る。次いで、各連結表現の尤度は、パスに基づく順位付け体系を使用して識別される（プロセス動作５０６）。既述したように、このパスに基づく順位付けは、連結表現において見出される個々のアトミック表現についてのスコアを結合することに関与し得る。

【００４８】

[0048] 候補変換プログラムの各々について確立される、前述された全体的な順位付けに関して、これらの順位付けを多様な手法で使用することが可能である。例えば、一実施形態において、候補変換プログラムの各々に関連付けられる順位付けは、所定の数の上位の候補変換プログラムを選択し、これら上位のプログラムを１つまたは複数の変換プログラムの前述された順位付けされたグループとして指定するために使用される。１つの実装において、所定の数は１であり、従って、最高位の候補変換プログラムのみが選択および指定されることに留意されたい。

【００４９】

１．６ 最高位の変換プログラムを使用して文字列変換を実行するための例示的なプロセス

[0049] 最高位の候補変換プログラムを選択することに関与する後者の実施形態に関して、これは、コンピューティング装置が文字列変換を実行する適用において使用され得る。より詳細には、図６を参照すると、１つの一般的な実装において、１つまたは複数のユー

ザー供給された入力－出力例から帰納的に合成された候補変換プログラムのセットが入力される（プロセス動作６００）。既述したように、候補変換プログラムの各々は、各ユーザー供給された入力例から各ユーザー供給された出力例によって提示される形式で出力文字列を生成する。

【００５０】

[0050] 入力されると、候補変換プログラムのセットは、順位付けされて、最高位の変換プログラムが識別される（プロセス動作６０２）。次いで、この最高位の変換プログラムが、ユーザー供給された入力文字列に適用されて、出力文字列が生成される（プロセス動作６０４）。

【００５１】

１．７ 例示的な変換シナリオ

[0051] このセクションでは、２つの例示的な変換シナリオが提供される。本明細書において説明される順位付け技法の実施形態は、これらのシナリオにおいて見出される少数の入力－出力例に基づいて、入力される各入力文字列から所望の出力文字列を生成することが可能な変換プログラムを確立するために採用され得る。

【００５２】

[0052] 第１の例示的なシナリオにおいて、スプレッドシートのユーザーは、入力列において一連の名前を有し、各名前の前に敬称「Ｍｒ．」を追加することを希望する。ユーザーは、図７に示される入力－出力例を提供した。

【００５３】

[0053] 課題は、出力文字列「Ｍｒ．Ｒｏｇｅｒ」中のどの部分文字列が定数文字列であり、どれが入力文字列「Ｒｏｇｅｒ」の部分文字列であるのかを決定することによって、所与の入力－出力例から所望の変換プログラムを学習することである。この場合には、出力部分文字列o_1［０．．０］≡Mが入力文字列中に存在しないため、定数文字列でなければならないことが推論され得る。しかし、出力部分文字列o_1［１．．１］≡ｒは、入力文字列中の２つの異なる部分文字列（i_1［０．．０］≡Ｒおよびi_1［４．．４］≡ｒ）に由来し得る。本明細書において説明される順位付け技法の実施形態は、全ての取り得る表現を学習して、（ｉ）部分文字列i_1［０．．０］を抽出し、「ｒ」から「Ｒ」を生成するための小文字演算を実行するための位置ペア表現と、（ｉｉ）部分文字列i_1［４．．４］および定数文字列表現「ｒ」を抽出するための位置ペア表現とを含む出力文字列中の部分文字列「ｒ」を算出する。入力文字列に存在しない出力文字列の部分文字列は、定数文字列であることが保証されるのに対して、入力文字列に存在しない部分文字列は、定数文字列または入力文字列の部分文字列のどちらかであり得る（経験は、これらが入力文字列に由来する可能性が高いことを示しているが）。例えば、出力部分文字列o_1［４．．８］≡Ｒｏｇｅｒは、定数文字列「Ｒｏｇｅｒ」よりも、入力文字列i_1に由来する可能性が高い。同様の議論を使用すると、出力文字列中の部分文字列o_1［１．．１］≡ｒも、入力文字列中の２つの位置に存在し、そこに由来する可能性が高い。しかし、この例において、プログラムの所望の振る舞いは定数文字列「Ｍｒ．」を各入力文字列の前に追加することであるため、「ｒ」は定数文字列である必要がある。１つの入力－出力例から所望の変換を学習するために、本明細書において説明される順位付け技法の実施形態は、定数文字列表現を出力部分文字列「ｒ」についての位置ペア表現よりも上位に順位付けする必要がある。「ｒ」を定数文字列として順位付けする際に役立ち得る特徴のうちの幾つかは、以下を含む。

ａ）部分文字列の長さ：部分文字列「ｒ」の長さは１であるため、部分文字列「ｒ」が入力部分文字列である可能性は低い。

ｂ）部分文字列の相対的な長さ：出力文字列と比較した部分文字列「ｒ」の相対的な長さも非常に小さく、すなわち、１／１０である。

ｃ）定数の隣接文字：「ｒ」の隣接文字「M」および「．」は、双方ともに定数表現である。

【００５４】

[0054] 第２の例示的なシナリオにおいて、スプレッドシートのユーザーは、列中に一連の住所を有し、この住所から都市名を抽出することを希望する。ユーザーは、図８に示される入力－出力例を提供した。

【００５５】

[0055] この場合には、本明細書において説明される順位付け技法の実施形態は、入力文字列「２４３　Ｆｌｙｅｒ　Ｄｒｉｖｅ，Ｃａｍｂｒｉｄｇｅ，ＭＡ　０２１４５」から部分文字列「Ｃａｍｂｒｉｄｇｅ」を抽出するために、１００を超える異なるＳｕｂＳｔｒ式を学習し得る。そのうちの幾つかは、以下の通りである。

ａ）p_1：３番目のアルファベットトークン文字列を抽出する。

ｂ）p_2：４番目の英数字のトークン文字列を抽出する。

ｃ）p_3：コンマおよび空白のトークンから成る１番目と２番目のトークンシーケンスの間の部分文字列を抽出する。

ｄ）p_4：３番目の大文字トークン（包括的）と左から２番目のコンマトークンとの間の部分文字列を抽出する。

【００５６】

[0056] 位置ペア式p_1およびp_2を学習することに関する問題は、入力文字列「５１２　Ｗｒｉｇｈｔ　Ａｖｅ，Ｌｏｓ　Ａｎｇｅｌｅｓ，ＣＡ　７８９１１」について、所望の出力ではない出力文字列「Ｌｏｓ」を生成することである。また、位置ペア式p_4は、入力文字列「６４　１２８ｔｈ　Ｓｔ，Ｓｅａｔｔｌｅ，ＷＡ　９８１０２」から、いかなる出力文字列も生成しない。他方、位置ペア式p_3は、位置ペア式の各々について所望の出力文字列（すなわち、それぞれCambridge、Los Angeles、SeattleおよびSan Francisco）を生成する。そのため、本明細書において説明される順位付け技法の実施形態は、位置ペア式p_3をその他の位置ペア式よりも上位に順位付けして、１つの入力－出力例から所望の出力文字列を生成する。

【００５７】

２．０　例示的な動作環境

[0057] 本明細書において説明される順位付け技法の実施形態は、多数のタイプの汎用または専用コンピューティングシステム環境または構成内で動作可能である。図９は、本明細書において説明されるような様々な実施形態および順位付け技法の実施形態の要素が実装され得る汎用コンピューターシステムの簡略化された例を例示する。図９において破線または点線によって表されるいかなる四角形も、簡略化されたコンピューティング装置の代替的な実施形態を表すこと、および、以下に説明されるように、これらの代替的な実施形態のうちのいずれかまたは全ては、本文書全体にわたって説明される他の代替的な実施形態と組み合わせて使用され得ることが留意されるべきである。

【００５８】

[0058] 例えば、図９は、簡略化されたコンピューティング装置１０を示す、一般的なシステム図を示す。そのようなコンピューティング装置は、典型的には、パーソナルコンピューター、サーバーコンピューター、ハンドヘルドコンピューティング装置、ラップトップまたはモバイルコンピューター、携帯電話およびＰＤＡなどの通信装置、マルチプロセッサシステム、マイクロプロセッサベースのシステム、セットトップボックス、プログラム可能な家庭用電化製品、ネットワークＰＣ、ミニコンピューター、メインフレームコンピューター、オーディオまたはビデオメディアプレーヤー等を含むが、これらに限定されない、少なくとも何らかの最低限の計算能力を有する装置において見出され得る。

【００５９】

[0059] 本明細書において説明される順位付け技法の実施形態を装置に実装させるために、装置は、基本的な計算演算を可能にするための充分な計算能力およびシステムメモリを有するべきである。特に、図９によって例示されるように、計算能力は、一般に、１つまたは複数の処理装置１２によって例示され、１つまたは複数のＧＰＵ１４も含み得、これらの一方または双方がシステムメモリ１６と通信する。一般的なコンピューティング装置の処理装置１２は、ＤＳＰ、ＶＬＩＷ、もしくは他のマイクロコントローラーなどの特殊

マイクロプロセッサであってもよく、または、マルチコアＣＰＵにおける特殊なＧＰＵベースのコアを含む、１つまたは複数の処理スコアを有する従来のＣＰＵとし得ることに留意されたい。

【００６０】

[0060] また、図９の簡略化されたコンピューティング装置は、例えば、通信インターフェイス１８などの他のコンポーネントも含み得る。図９の簡略化されたコンピューティング装置は、１つまたは複数の従来のコンピューター入力装置２０（例えば、ポインティング装置、キーボード、オーディオ入力装置、ビデオ入力装置、触覚入力装置、有線または無線データ送信を受信するための装置等）も含み得る。図９の簡略化されたコンピューティング装置は、例えば、１つまたは複数の従来のディスプレイ装置２４および他のコンピューター出力装置２２（例えば、オーディオ出力装置、ビデオ出力装置、有線または無線データ送信を送信するための装置等）などの、他の任意選択的なコンポーネントも含み得る。汎用コンピューターのための典型的な通信インターフェイス１８、入力装置２０、出力装置２２、および記憶装置２６は、当業者には周知であり、本明細書において詳細に説明されないことに留意されたい。

【００６１】

[0061] 図９の簡略化されたコンピューティング装置は、多様なコンピューター読取可能な媒体も含み得る。コンピューター読取可能な媒体は、コンピューター１０によって記憶装置２６を介してアクセスされ得る任意の利用可能な媒体とすることができ、コンピューター読取可能な命令もしくはコンピューター実行可能な命令、データ構造、プログラムモジュール、または他のデータなどの情報の記憶のために、取り外し可能２８および／または取り外し不可能３０である揮発性媒体および不揮発性媒体の双方を含む。コンピューター読取可能な媒体は、コンピューター記憶媒体および通信媒体を含み得る。コンピューター記憶媒体は、ＤＶＤ、ＣＤ、フロッピーディスク、テープドライブ、ハードディスクドライブ、光学式ドライブ、ソリッドステートメモリドライブ、ＲＡＭ、ＲＯＭ、ＥＥＰＲＯＭ、フラッシュメモリもしくは他のメモリ技術、磁気カセット、磁気テープ、磁気ディスク記憶装置もしくは他の磁気記憶装置、または所望の情報を記憶するために使用されることができ、かつ、１つまたは複数のコンピューティング装置によってアクセスされることができる任意の他の装置などの、有形のコンピューター読取可能なまたは機械読取可能な媒体または記憶装置をいう。

【００６２】

[0062] コンピューター読取可能なまたはコンピューター実行可能な命令、データ構造、プログラムモジュール等などの情報の保持は、１つもしくは複数の変調データ信号もしくは搬送波を符号化するための多様な前述の通信媒体のいずれか、または、他の伝送機構もしくは通信プロトコルを使用することによっても達成され得、任意の有線もしくは無線の情報配信機構を含む。「変調データ信号」または「搬送波」という用語は、一般的に、その特性のうちの１つまたは複数が情報を信号に符号化するような手法で設定または変更された信号をいうことに留意されたい。例えば、通信媒体は、１つまたは複数の変調データ信号を搬送する有線ネットワークまたは直接有線接続などの有線媒体、ならびに音響、ＲＦ、赤外線、レーザー、および１つまたは複数の変調データ信号または搬送波を送信および／または受信するための他の無線媒体などの無線媒体を含む。上記のうちの任意のものの組み合わせも、通信媒体の範囲内に含まれるべきである。

【００６３】

[0063] さらに、本明細書において説明される様々な順位付け技法の実施形態のうちの一部または全部を具現化するソフトウェア、プログラムおよび／もしくはコンピュータープログラム製品、またはこれらの一部は、コンピューターまたは機械読取可能な媒体または記憶装置および通信媒体の任意の所望の組み合わせから、コンピューター実行可能な命令または他のデータ構造の形式において、記憶され、受信され、送信され、または読み出され得る。

【００６４】

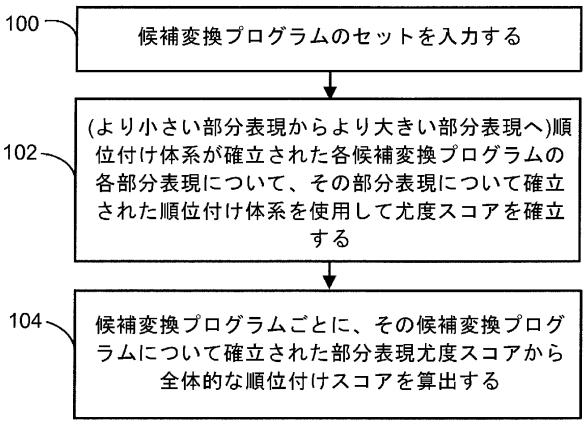
[0064] 最後に、本明細書において説明される順位付け技法の実施形態は、コンピューティング装置によって実行される、プログラムモジュールなどのコンピューター実行可能な命令の一般的なコンテキストにおいてさらに説明され得る。一般に、プログラムモジュールは、特定のタスクを実行し、または特定の抽象データ型を実装するルーチン、プログラム、オブジェクト、コンポーネント、データ構造等を含む。本明細書において説明される実施形態は、１つもしくは複数の通信ネットワークを通じてリンクされる、タスクが１つもしくは複数の遠隔処理装置によって実行される分散コンピューティング環境において、または１つもしくは複数の装置のクラウド内でも実施され得る。分散コンピューティング環境において、プログラムモジュールは、媒体記憶装置を含むローカルコンピューターおよび遠隔コンピューターの記憶媒体の双方において位置し得る。またさらに、前述された命令は、一部または全部が、ハードウェア論理回路として実装されてもよく、ハードウェア論理回路は、プロセッサを含んでも、または含まなくてもよい。

【００６５】

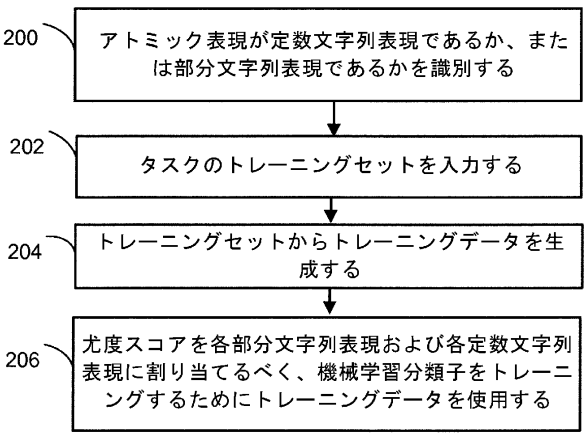
３．０ 他の実施形態

[0065] 本説明全体にわたる前述された実施形態のいずれかまたは全部は、任意の所望の組み合わせにおいて使用されて、付加的な混合実施形態を形成し得ることに留意されたい。また、主題は、構造的な特徴および／または方法論的な動作に固有の言語において説明されてきたが、添付の特許請求の範囲において定義される主題は、必ずしも上述された具体的な特徴または動作に限定されないことが理解されるべきである。むしろ、上述された具体的な特徴および動作は、請求項を実装する例示的な形式として開示される。

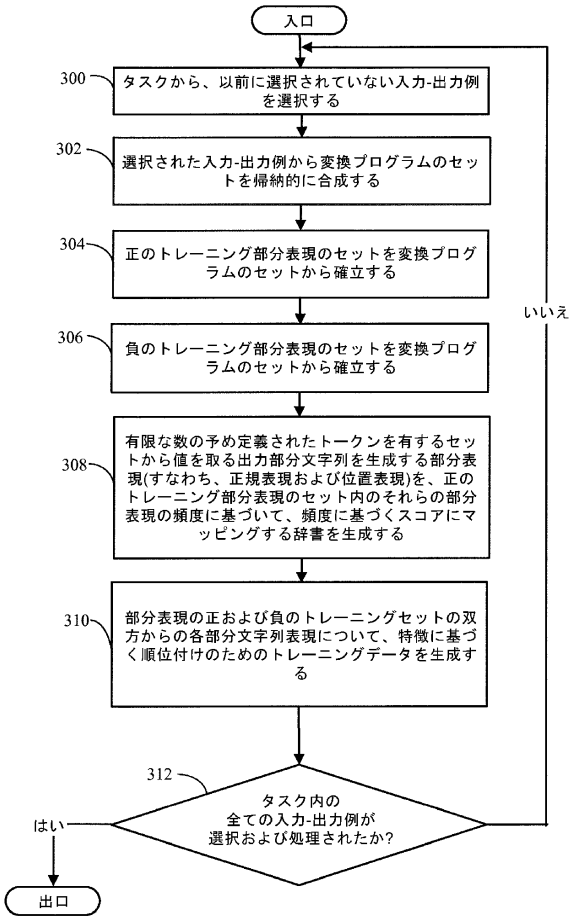
【図１】



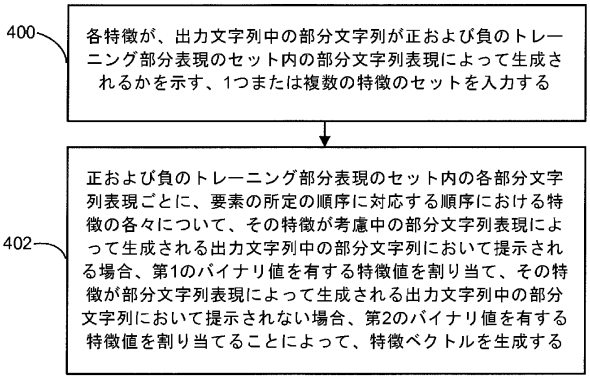
【図２】



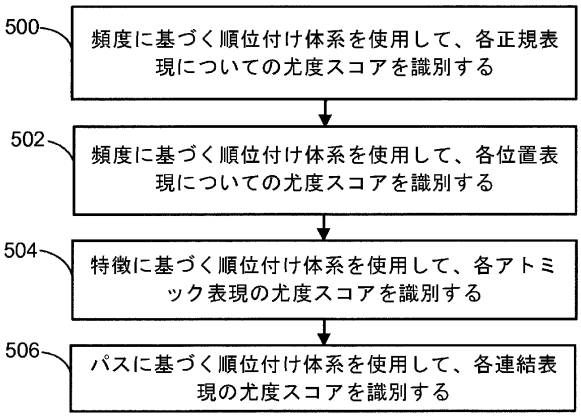
【図３】



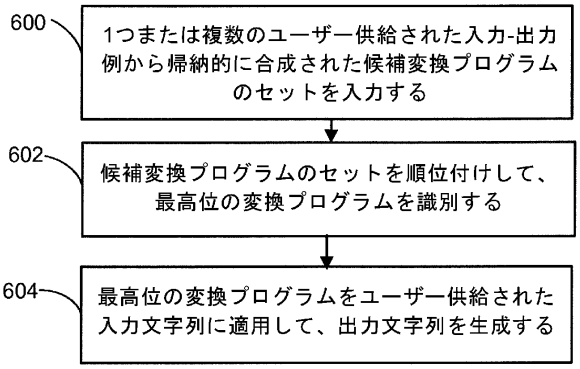
【図４】



【図５】



【図６】



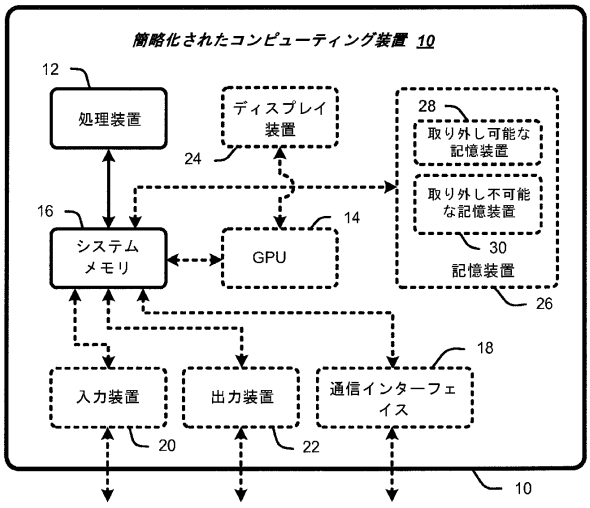
【図７】

	入力v_1	出力
1	Roger	Mr. Roger
2	Simon	
3	Benjamin	
4	John	

【図８】

	入力v_1	出力
1	243 Flyer Drive, Cambridge, MA 02145	Cambridge
2	512 Wright Ave, Los Angeles, CA 78911	
3	64 128th St, Seattle, WA 98102	
4	560 Hearst Ave, San Francisco, CA 94129	

【図９】



フロントページの続き

(74)代理人 100108213
弁理士 阿部 豊隆
(74)代理人 100140431
弁理士 大石 幸雄
(72)発明者 ガルワニ, スミット
アメリカ合衆国, ワシントン州 ９８０５２－６３９９, レッドモンド, ワン マイクロソフト ウェイ, マイクロソフト コーポレーション内, エルシーエー － インターナショナル パテンツ
(72)発明者 シン, リシャブ
アメリカ合衆国, ワシントン州 ９８０５２－６３９９, レッドモンド, ワン マイクロソフト ウェイ, マイクロソフト コーポレーション内, エルシーエー － インターナショナル パテンツ

審査官 成瀬 博之

(56)参考文献 米国特許出願公開第２０１１／０３０２５５３（ＵＳ, Ａ1）
米国特許出願公開第２０１２／０１９２０５１（ＵＳ, Ａ1）
特開平０７－３１９６８２（ＪＰ, Ａ）

(58)調査した分野(Int.Cl., ＤＢ名)
Ｇ０６Ｆ １７／２０－１７／２８
Ｇ０６Ｆ ８／６５
Ｇ０６Ｎ ９９／００